PSYCHOMETRIKA—VOL. 74, NO. 4, 603–618
DECEMBER 2009
DOI: 10.1007/S11336-009-9122-3

A TRUNCATED-PROBIT ITEM RESPONSE MODEL FOR ESTIMATING PSYCHOPHYSICAL THRESHOLDS

RICHARD D. MOREY

UNIVERSITY OF GRONINGEN

JEFFREY N. ROUDER AND PAUL L. SPECKMAN

UNIVERSITY OF MISSOURI

Human abilities in perceptual domains have conventionally been described with reference to a threshold that may be defined as the maximum amount of stimulation which leads to baseline performance. Traditional psychometric links, such as the probit, logit, and t, are incompatible with a threshold as there are no true scores corresponding to baseline performance. We introduce a truncated probit link for modeling thresholds and develop a two-parameter IRT model based on this link. The model is Bayesian and analysis is performed with MCMC sampling. Through simulation, we show that the model provides for accurate measurement of performance with thresholds. The model is applied to a digit-classification experiment in which digits are briefly flashed and then subsequently masked. Using parameter estimates from the model, individuals' thresholds for flashed-digit discrimination is estimated.

Key words: IRT, item response theory, threshold, thresholds, psychometrics, psychophysics, Bayesian hierarchical models, MAC, mass at chance.

Psychometrics refers to the measurement of human abilities and characteristics. Not surprisingly, psychometric theories have been developed in many contexts including education and psychopathology. One area that does not seem to overlap with psychometrics is psychophysics, which is the measurement of perception. Historically, one of the main constructs in psychophysics was that of a threshold or limen on stimulus intensity. If a stimulus intensity is below the threshold, the stimulus cannot be perceived. Conversely, if stimulus intensity is above the threshold, the stimulus is perceived to some degree. Throughout, we use stimulus intensity generically to refer to any stimulus strength dimension, including brightness, loudness, duration, etc.

The conventional psychophysical approach to measuring thresholds is adaptive staircasing (Taylor & Creelman, 1967; Watson & Pelli, 1983). Staircase procedures are designed to find the stimulus intensity such that performance is at some specified level. Consider the case where stimulus A and B are presented equally often and the participant decides on each trial whether A or B was presented. Accuracy in this task ranges from a baseline of 0.5, indicating chance performance, to ceiling performance of 1.0. Commonly specified levels for thresholds in adaptive psychophysical procedures are 0.75 and 0.707, the long-run convergence probabilities of specific adaptive staircase methods.

In some domains and for some questions, however, these specified levels are not appropriate. Consider the example of subliminal priming. In subliminal priming, the claim is that stimuli which are identified at a level no better than the chance baseline nonetheless affect subsequent

This research is part of the first author's Ph.D. thesis from the University of Missouri. We thank Mike Pratte and Andrew Kent for help in running the reported experiment. This research is supported by NSF grant SES-0351523 and NIMH grant R01-MH071418 and an Adeline Hoffman Fellowship from the University of Missouri.

Requests for reprints should be sent to Richard D. Morey, DPMG, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands. E-mail: r.d.morey@rug.nl

behavior (e.g., Abrams, Klinger, & Greenwald, 2002; Dehaene, Naccache, Le Clech, Koechlin, Mueller, & Dehaene-Lambertz, 1998; Eimer & Schlaghecken, 2002; Snodgrass, Bernat, & Shevrin 2004; Vorberg, Mattler, Heinecke, Schmidt, & Schwarzbach, 2003). To test for subliminal priming, it is critical that performance be at chance rather than above chance (Reingold & Merikle, 1988; Rouder, Morey, Speckman, & Pratte, 2007). There is no current staircase procedure, however, that converges to a chance level of performance. In fact, simpler methods, such as confidence intervals, fail as well. Because the size of confidence intervals is inversely proportional to $\sqrt{N}$, very large sample sizes are required to differentiate baseline accuracy from just-above baseline accuracy. Consider an experimenter who wants to differentiate baseline performance of 0.5 from a just-above baseline performance of 0.52. If the observed proportion correct was exactly 0.50, the experimenter would need a confidence interval of width less than 0.04 to rule out 0.52. At $\alpha = 0.05$, about 2,400 trials would be required.

In this paper, we describe how item response theory (IRT, Lord & Novick, 1968) may be adapted to mitigate these difficulties. The model may be used for two complementary purposes. First, it may be used to classify a participant's performance to a given stimulus intensity as either at chance or above chance. Second, it may be used to identify the stimulus intensity that corresponds to the transition from at-chance to above-chance performance. To model this transition, we use a truncated-probit link between latent scores and performance. Model development includes that of a new Metropolis step to decorrelate MCMC chains. After showing that the model performs well in simulation, we apply the model to measure thresholds in a simple perceptual task. Although our application is in perception, the notion of thresholds may be applicable in other measurement domains such as education and psychopathology.

1. Mass at Chance Item Response Theory Models

IRT models provide accurate estimates of abilities and item difficulties in reasonable sample sizes by pooling information within a hierarchical structure. We use the same strategy here and adapt IRT for the measurement of thresholds by pooling across individuals and intensity levels. We develop models for the case in which the experimenter presents either Stimulus A or Stimulus B at a number of intensity levels. The participant is asked to identify the stimulus as A or B, and accuracy, which ranges form .5 to 1.0, serves as the dependent measure. Each stimulus is presented equally often at each intensity level. Although we deal solely with this case for simplicity of exposition, generalization beyond two choices is straightforward.

Consider the following one-parameter logistic (1PL) model of performance in a psychophysical task. We refer to each intensity level of the stimulus as an item. Let y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, denote the number of correct responses in N_{ij} attempts for the ith participant observing the jth item:

$$y_{ij} \overset{\text{indep.}}{\sim} \text{Binomial}(N_{ij}, p_{ij}),$$

where p_{ij} is the true probability of correct response. In 1PL,

$$p_{ij} = \frac{1}{2} + \frac{1}{2(1 + e^{-(\alpha_i - \beta_j)})}, \tag{1}$$

where α_i and β_j denote the ability and difficulty of the ith participant and jth item, respectively. Participant abilities are zero-centered normal random variables with common variance. This model may be analyzed with a variety of methods (i.e., Andersen, 1973; Rasch, 1960; Swaminathan & Gifford, 1982).

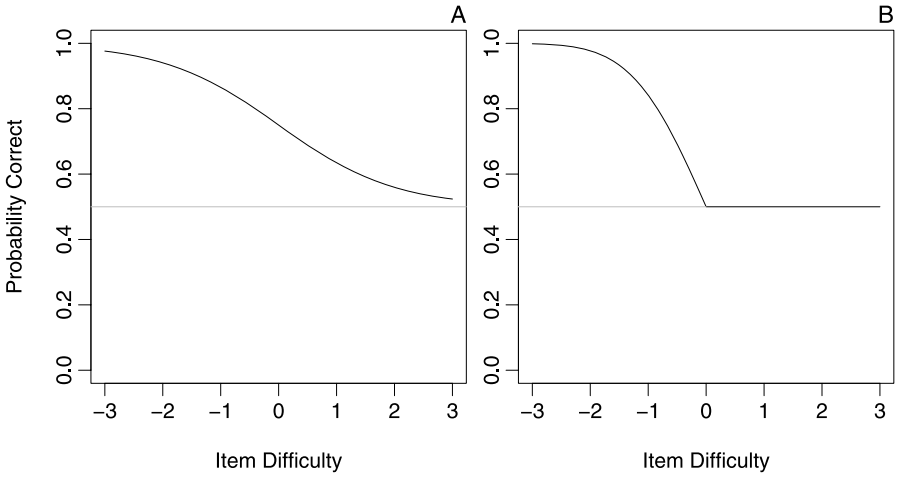

FIGURE 1.
A. Person-response function for the average person in 1PL. **B**. The same for the 1P-MAC model.

The model in (1) differs from the conventional 1PL in that the minimum performance is 0.5 rather than 0. Models with baseline parameters other than 0 are often considered members of the 3PL family, but we call the model a 1PL model. In naming the models throughout, we follow the naming convention in Verhelst and Glas (1995), who distinguished between fixed and free parameters. Our baseline is fixed naturally to 0.5 by considerations of the design. The value 0.5 plays an analogous role to 0 in conventional 1PL in that it is not estimated. Hence, from a structural view, the effects of items and people are solely to shift true scores, as they are in the conventional 1PL model. To highlight this structural equivalence, we consider the model a member of the 1PL family. We follow throughout the Verhelst and Glas convention that a model is named with reference to free rather than fixed parameters.

One drawback to the 1PL model in (1) is that it does not model a transition between chance performance and above chance performance; it disallows thresholds, as we have defined them.[1] An individual's threshold in this case is the minimum value of β_j necessary for $p_{ij} = 0.5$. However, for all finite values of α_i and β_j, $p_{ij} > .5$. Figure 1A shows this fact. The line is a person-response curve for the average person ($\alpha_i = 0$). As difficulty increases, accuracy approaches the baseline but never attains it. This incompatibility with thresholds holds for all links based on CDFs with support on $(-\infty, \infty)$, including the logit, probit, and t links.

Rouder et al. (2007) proposed a truncated-probit link to model the transition in performance from baseline to above-baseline levels when multiple participants responded repeatedly to a single item:

$$p(\alpha, \beta) = \begin{cases} \Phi(\alpha_i - \beta), & \alpha_i > \beta, \\ .5, & \beta \geq \alpha_i, \end{cases} \tag{2}$$

where Φ is the CDF of the standard normal distribution. If the person's ability is greater than the item's difficulty, then performance is above baseline; conversely, if the item's difficulty is greater than the person's ability, then performance is at baseline. A person-response curve for

[1] The term *threshold* has a unique definition in different areas of psychology, and IRT is no exception. In IRT models, *threshold* refers to the additive ability parameter (denoted α_i throughout). In order to mitigate any possible confusion, we use the term *ability* or *true score* when refering to latent performance variables and the term *threshold* when referring to a level of intensity where performance transitions from an at-chance level to an above-chance level.

an average person ($\alpha_i = 0$) is shown in Figure 1B. For all positive difficulties, performance is at baseline. For all negative difficulties, performance follows a probit. Because there is nonzero probability of baseline performance, Rouder et al. called this model *Mass at Chance* (MAC). We refer to the one-parameter MAC model as 1P-MAC and the two parameter extension, presented subsequently, as 2P-MAC.

Rouder et al. developed the MAC model for a single item with the goal of simply classifying participants' accuracies as either above or at baseline. Morey, Rouder, and Speckman (2008) developed 1P-MAC for multiple items and used it to both classify participant-by-item combinations as well as estimate individuals' thresholds. In this paper, we develop a 2P-MAC model for classification and threshold estimation. Our motivation for this extension is a noticeable misspecification of the Morey et al. (2008) one-parameter model in the analysis of data from a digit identification task. We show this misspecification in the next section. Following that, we present a 2P-MAC model and show how it accounts for Morey et al. (2008) data set.

2. Misspecification of the One-Parameter Mass at Chance Model

The 1P-MAC model is based on an additivity assumption borrowed from Rasch IRT models. The assumption implies that all participants improve at the same rate as item difficulty is decreased. Morey et al. (2008) show through simulation that the 1P-MAC model provides for accurate estimates of thresholds in reasonably-sized samples when additivity holds. There are two critical questions regarding this assumption: first, what are the consequences of violating this assumption; and, second, is it violated in empirical data? If the consequences are severe and the assumption is violated, the model must be expanded for accurate analysis.

We first explore the effects of assuming additivity when it does not hold. The solid lines in Figure 2 show two hypothetical participants with the same threshold but different rates of improvement. These curves violate the additivity in the 1P-MAC model as they are not shifts of one another. To account for the difference in performance between these participants, the 1P-MAC model estimates the thresholds as being different. Thus, participant A's threshold is underestimated, and participant B's threshold is overestimated (Figure 2, dashed lines). Because the estimation of thresholds is one of the primary purposes of MAC models, violations of this assumption are serious challenges to the usefulness of the 1P-MAC model.

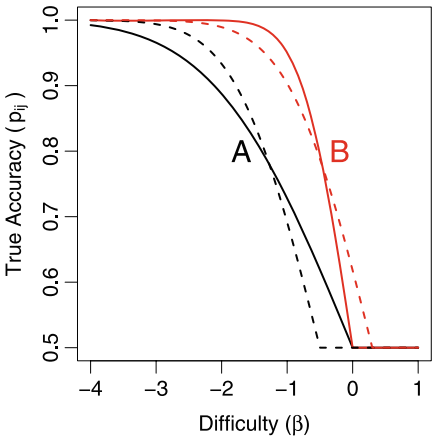

FIGURE 2.
The effect of differing rates of improvement in the 1P-MAC model. *Solid lines A* and *B* represent true latent scores at different item difficulties. *Dashed lines* represent the hypothetical 1P-MAC model fit.

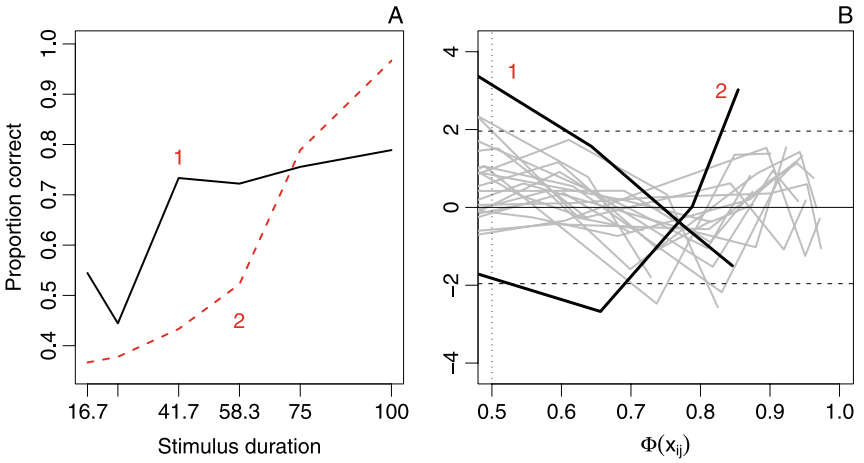

FIGURE 3.
A. Accuracy from two selected participants in Morey et al. (2008) demonstrate the plausibility of participant discriminability. **B**. Analysis of the 1P-MAC model shows that a few participants have different discriminabilities than the others.

With the consequences of misspecification established, it is reasonable to ask whether additivity appears to hold in data. Morey et al. (2008) provide a suitable data set. They asked 22 participants to classify single digits as being less-than five or greater-than five. The digits were flashed briefly and followed by a punctuation mark that served as a pattern mask (Breitmeyer, 1984). If the flashes are sufficiently brief, performance is greatly degraded and near the chance level of 0.5. Morey et al. employed six duration levels ranging from 16.7 ms to 100 ms and asked each participant to classify 90 digits at each duration level.

Figure 3A shows accuracies for two selected participants as a function of duration. The solid line, from a selected participant, shows performance that rises above chance at brief exposures, and improves slowly thereafter. The dotted line, from a different participant, shows performance that rises above chance only after long exposures, and improves quickly thereafter. This behavior violates the additivity assumption in 1P-MAC. The effects may also be seen in the standardized residuals from the 1P-MAC model analysis (Panel B). If the model is correct, then these standardized residuals should be distributed approximately as a standard normal (we discuss the computation of these residuals subsequently). Each line shows the residuals for a participant. For most participants, denoted with gray lines, the residuals are fairly small and without systematic trends. For a few participants, however, the residuals are large and patterned. Participant 1, for example, gains at a rate slower than most; Participant 2 gains at a rate faster than most.

The strong possibility of different rates of improvement in the Morey et al. (2008) data set motivates the development of the following 2P-MAC model. For psychophysical applications, we develop the model in which each participant has his or her unique discriminability parameter. This development is in contrast to standard two-parameter models in which items rather than people have discriminability parameters. In psychophysical settings, in contrast to other testing situations, items typically vary on a single dimension and should be modeled accordingly. People, on the other hand, may vary on more than one dimension. The following development is sufficiently general that it is straightforward to adapt the model for item discriminability rather than participant discriminability. In this regard, the presented 2P-MAC model may be used as an alternative to two-parameter logistic or probit IRT models.

3. The 2P-MAC Model

3.1. Model Specification

The 2P-MAC model is given as

$$p_{ij} = \begin{cases} \Phi(\theta_i(\alpha_i - \beta_j)), & \alpha_i > \beta_j, \\ .5, & \beta_j \geq \alpha_i. \end{cases} \tag{3}$$

In the following development, it is convenient to define a latent true score, x_{ij} for each participant-by-item combination: $x_{ij} = \theta_i(\alpha_i - \beta_j)$.

The difference between the 1P-MAC and the 2P-MAC model is the inclusion of the θ_i parameters. Although increasing the number of parameters will typically improve model fit, we are not adding the parameters to combat a generic lack-of-fit. Rather, these parameters are necessary to account for a specific, predictable type of misfit: different rates of participant improvement. The θ_i parameters are interpreted as indexing these individualized rates of improvement. In the subsequent analysis of the Morey et al. (2008) data, we show that the improvement in model fit from 1P-MAC to 2P-MAC is more than enough to make up for the loss in model parsimony.

The model needs an additional constraint. In conventional two-parameter IRT models with item discriminability, this constraint is placed on the variance of participant effects. In the model in (3), in which discriminability varies by individuals rather than items, it is convenient to place the constraint on item difficulties rather than participant effects:

$$\beta_j \overset{iid}{\sim} \text{Normal}(0, 1). \tag{4}$$

We show subsequently that this constraint works well in simulation and application.

The model is analyzed in the Bayesian framework with Markov chain Monte Carlo sampling (Geman & Geman, 1984; Gelfand & Smith, 1990, Gelman, Carlin, Stern, & Rubin, 2004). Consequently, priors are needed for parameters. The following hierarchical priors are convenient for participant effects:

$$\begin{aligned} \alpha_i &\overset{iid}{\sim} \text{Normal}\left(0, \sigma_\alpha^2\right), \\ \theta_i &\overset{iid}{\sim} \text{TN}_{\Re^+}\left(0, \sigma_\theta^2\right), \\ \sigma_\alpha^2 &\sim \text{Inverse Gamma}(a_1, b_1), \\ \sigma_\theta^2 &\sim \text{Inverse Gamma}(a_2, b_2), \end{aligned}$$

where TN_A denotes a normal distribution truncated to an interval A. The inverse gamma prior for variance parameters has pdf

$$f(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\left\{-\frac{b}{x}\right\}, \quad x > 0,$$

for shape parameter a and scale parameter b. Values of a_1, b_1, a_2, and b_2 must be specified before analysis. We find $a_1 = a_2 = 2$, and $b_1 = b_2 = 1$ to work well in simulation and application. These prior distributions are standard in Bayesian analysis, with the possible exception of the truncated normal prior on the discriminability parameters θ_i. We chose this prior over other positive priors such as the gamma or lognormal for computational convenience. The truncated normal on discriminability is conjugate in this setting, that is, the conditional posterior of θ_i is also a truncated normal. Sampling from a truncated normal is convenient and may be done without more computationally expensive Metropolis steps.

3.2. *Full Conditional Distributions*

Derivation of the full conditional distributions is aided by the addition of the following latent variables (Albert & Chib, 1993). Let y_{ijk} be the dichotomous response for the ith participant observing the jth item for the kth replicate, $k = 1, \ldots, N_{ij}$. Let w_{ijk} denote latent data such that $y_{ijk} = 1 \iff w_{ijk} > 0$. If

$$w_{ijk} \overset{indep}{\sim} \text{Normal}\left(\theta_i(\alpha_i - \beta_j) \vee 0, 1\right)$$

then

$$P(w_{ijk} > 0) = P(y_{ijk} = 1),$$

where $a \vee b$ denotes $\max(a, b)$.

It is more convenient to place the hierarchical model on latent data $\mathbf{w}$ than on observed data $\mathbf{y}$. We use bold-face notation throughout to represent collections and vectors of data or parameters. The following facts describe all the full conditional distributions used in MCMC analysis, including those for the latent data $\mathbf{w}$.

Fact 1. The full conditional posteriors of w_{ijk}, σ_α^2, and σ_θ^2 are

$$w_{ijk} \mid \cdot \sim \begin{cases} \text{TN}_{\Re+}(x_{ij} \vee 0, 1), & y_{ijk} = 1, \\ \text{TN}_{\Re-}(x_{ij} \vee 0, 1), & y_{ijk} = 0, \end{cases}$$

$$\sigma_\alpha^2 \mid \cdot \sim \text{IG}\left(a_1 + \frac{I}{2}, b_2 + \frac{1}{2}\sum_{i=1}^{I} \alpha_i^2\right),$$

$$\sigma_\theta^2 \mid \cdot \sim \text{IG}\left(a_2 + \frac{I}{2}, b_2 + \frac{1}{2}\sum_{i=1}^{I} \theta_i^2\right),$$

where $\cdot$ refers to all other parameters and data.

Fact 2. The full conditional posteriors of α_i are independent for given i, $\boldsymbol{\beta}$, θ_i, σ_α^2, and $\mathbf{w}$. It is convenient to define the following sets. Let $A_0 = (-\infty, \beta_{(1)})$, $A_1 = (\beta_{(1)}, \beta_{(2)})$, $\ldots$, $A_J = (\beta_{(J)}, \infty)$, where $\beta_{(1)} < \cdots < \beta_{(J)}$ are the order statistics of the β_j. Let

$$J_\ell = \{j : \beta_j \leq \beta_{(\ell)}\} \tag{5}$$

and define

$$s_{\alpha_i \ell} = \begin{cases} (\frac{1}{\sigma_\alpha^2} + \sum_{j \in J_\ell} N_{ij}\theta_i^2)^{-1}, & 0 < \ell \leq J, \\ \sigma_\alpha^2, & \ell = 0, \end{cases} \tag{6}$$

$$m_{\alpha_i \ell} = \begin{cases} s_{\alpha_i \ell} \sum_{j \in J_\ell} (\theta_i \sum_{k=1}^{N_{ij}} w_{ijk} + N_{ij}\theta_i^2 \beta_j), & 0 < \ell \leq J, \\ 0, & \ell = 0, \end{cases} \tag{7}$$

$$h_{\alpha_i \ell} = \begin{cases} \sum_{j \in J_\ell} (N_{ij}\theta_i^2\beta_j^2 + 2\theta_i\beta_j \sum_{k=1}^{N_{ij}} w_{ijk}), & 0 < \ell \leq J, \\ 0, & \ell = 0. \end{cases} \tag{8}$$

Then

$$[\alpha_i \mid \cdot] \propto \sum_{\ell=0}^{J} \exp\left\{-\frac{1}{2}\left(h_{\alpha_i\ell} - \frac{(m_{\alpha_i\ell})^2}{s_{\alpha_i\ell}}\right)\right\} \times \exp\left\{-\frac{(\alpha_i - m_{\alpha_i\ell})^2}{2s_{\alpha_i\ell}}\right\} I_{(\alpha_i \in A_\ell)}. \tag{9}$$

Following Bayesian convention, $[\alpha_i \mid \cdot]$ denotes the density of α_i given all other parameters and data. The right-hand side of (9) is proportional to the density of a mixture of truncated normals. Set $\beta_{(J+1)} = \infty$ and $\beta_{(0)} = -\infty$. Let

$$q_{\alpha_i\ell} = \exp\left\{-\frac{1}{2}\left(h_{\alpha_i\ell} - \frac{(m_{\alpha_i\ell})^2}{s_{\alpha_i\ell}}\right)\right\} \times \sqrt{2\pi s_{\alpha_i\ell}}\left[\Phi\left(\frac{\beta_{(\ell+1)} - m_{\alpha_i\ell}}{\sqrt{s_{\alpha_i\ell}}}\right) - \Phi\left(\frac{\beta_{(\ell)} - m_{\alpha_i\ell}}{\sqrt{s_{\alpha_i\ell}}}\right)\right] \tag{10}$$

for $\ell = 0, \ldots, J$, and let $p_{\alpha_i\ell} = q_{\alpha_i\ell} / \sum_{\ell=0}^{J} q_{\alpha_i\ell}$, $\ell = 0, \ldots, J$. Then the full conditional distribution $[\alpha_i \mid \cdot]$ is the mixture of truncated normal distributions

$$\alpha_i \mid \cdot \sim \sum_{\ell=0}^{J} p_{\alpha_i\ell} \mathrm{TN}_{A_\ell}(m_{\alpha_i\ell}, s_{\alpha_i\ell}). \tag{11}$$

Fact 3. The full conditional posterior distributions of the β_j are independent for given $j, \boldsymbol{\alpha}, \boldsymbol{\theta}$, and $\mathbf{w}$. To define them, let $M_0 = (\alpha_{(J)}, \infty)$, $M_1 = (\alpha_{(J-1)}, \alpha_{(J)})$, $\ldots$, $M_I = (-\infty, \alpha_{(1)})$, where $\alpha_{(1)}, \ldots, \alpha_{(I)}$ are the order statistics of the α_i. Let

$$I_\ell = \{i : \alpha_i > \alpha_{(I-\ell)}\}$$

and define

$$s_{\beta_j\ell} = \begin{cases} (\frac{1}{\sigma_\beta^2} + \sum_{i \in I_\ell} N_{ij}\theta_i^2)^{-1}, & 0 < \ell \le I, \\ \sigma_\beta^2, & \ell = 0, \end{cases}$$

$$m_{\beta_j\ell} = \begin{cases} s_{\beta_j\ell}(-\sum_{i \in I_\ell}(\theta_i \sum_{k=1}^{N_{ij}} w_{ijk} - N_{ij}\theta_i^2\alpha_i)), & 0 < \ell \le I, \\ 0, & \ell = 0, \end{cases}$$

$$h_{\beta_j\ell} = \begin{cases} \sum_{i \in I_\ell}(N_{ij}\theta_i^2\alpha_i^2 - 2\theta_i\alpha_i \sum_{k=1}^{N_{ij}} w_{ijk}), & 0 < \ell \le I, \\ 0, & \ell = 0. \end{cases}$$

Then

$$[\beta_j \mid \cdot] \propto \sum_{\ell=0}^{I} \exp\left\{-\frac{1}{2}\left(h_{\beta_j\ell} - \frac{(m_{\beta_j\ell})^2}{s_{\beta_j\ell}}\right)\right\} \times \exp\left\{-\frac{(\beta_j - m_{\beta_j\ell})^2}{2s_{\beta_j\ell}}\right\} I_{(\beta_j \in M_\ell)}.$$

The right-hand side is proportional to the density of a mixture of truncated normal distributions. Set $\alpha_{(I+1)} = \infty$ and $\alpha_{(0)} = -\infty$. Let

$$q_{\beta_j \ell} = \exp\left\{-\frac{1}{2}\left(h_{\beta_j \ell} - \frac{(m_{\beta_j \ell})^2}{s_{\beta_j \ell}}\right)\right\} \times \sqrt{2\pi s_{\beta_j \ell}}\left[\Phi\left(\frac{\alpha_{(I-\ell+1)} - m_{\beta_j \ell}}{\sqrt{s_{\beta_j \ell}}}\right) - \Phi\left(\frac{\alpha_{(I-\ell)} - m_{\beta_j \ell}}{\sqrt{s_{\beta_j \ell}}}\right)\right]$$

for $\ell = 0, \ldots, I$, and let $p_{\beta_j \ell} = q_{\beta_j \ell} / \sum_{\ell=0}^{I} q_{\beta_j \ell}$, $\ell = 0, \ldots, I$. Then the full conditional density $[\beta_j \mid \cdot]$ is the mixture of truncated normal distributions

$$\beta_j \mid \cdot \sim \sum_{\ell=0}^{I} p_{\beta_j \ell} \mathrm{TN}_{M_\ell}(m_{\beta_j \ell}, s_{\beta_j \ell}). \tag{12}$$

Fact 4. The full conditional posterior distributions of θ_i are independent for given α_i, $\boldsymbol{\beta}$, and $\mathbf{w}$. Let

$$J^* = \{j : \beta_j < \alpha_i\},$$

$$s_{\theta_i} = \left(\frac{1}{\sigma_\theta^2} + \sum_{j \in J^*} N_{ij}(\alpha_i - \beta_j)^2\right)^{-1},$$

$$m_{\theta_i} = s_{\theta_i} \sum_{j \in J^*} (\alpha_i - \beta_j) \sum_{k=1}^{N_{ij}} w_{ijk}.$$

Then

$$\theta_i \mid \cdot \sim \mathrm{TN}_{\Re^+}(m_{\theta_i}, s_{\theta_i}). \tag{13}$$

Proofs of these four facts follow from the proofs in Morey et al. (2008) and are omitted.

3.3. Model Analysis

The full conditional posterior distributions in Facts 1 through 4 are inverse gamma, truncated normal, and mixtures of truncated normal distributions. Morey et al. (2008) provide numerically precise, efficient sampling routines for these distributions. Marginal posterior distributions may be obtained through Gibbs sampling (Gelfand & Smith, 1990; see Rouder & Lu, 2005 for a tutorial). Although MCMC methods, such as Gibbs sampling, are guaranteed to converge to the joint posterior distribution under fairly weak conditions that are met here (Tierney, 1994), convergence may be slow due to substantial correlation from iteration to iteration. If this correlation is especially pronounced, the convergence may take hours, days, or longer.

Unfortunately, the Gibbs sampler for the 2P-MAC model produces highly autocorrelated chains. Figure 4A shows a chain for a typical item parameter, β_4, from the analysis of the Morey et al. (2008) data set. The high degree of autocorrelation is evidenced by the slowly wandering nature of the chain. The source of the autocorrelation is straightforward to diagnose. As shown in (3), the likelihood of the model given the parameters is a function of $\theta_i(\alpha_i - \beta_j)$. The likelihood is invariant to the inclusion of additive and multiplicative constants as follows: Let $\alpha_i^* = v(\alpha_i + z)$, $\beta_j^* = v(\beta_j + z)$ and $\theta_i^* = \theta_i / v$. Then the likelihood is the same, regardless of what the values of z and v are. Hence, values of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ cannot be identified from the likelihood

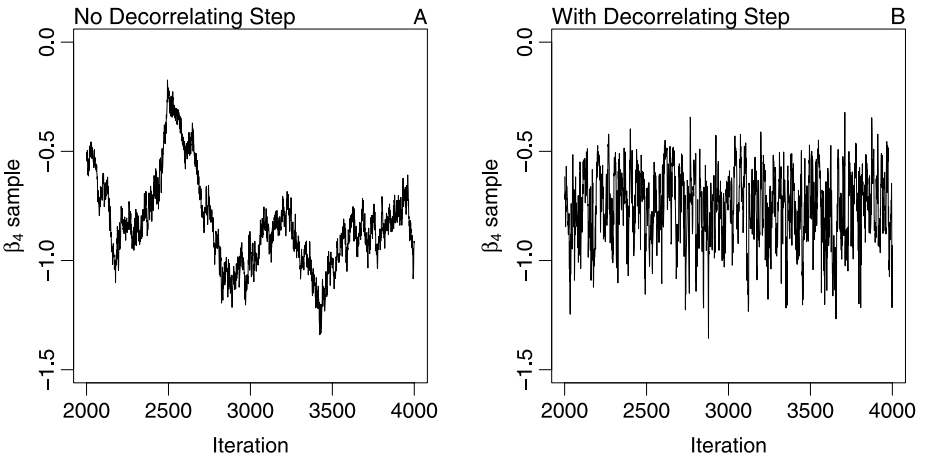

FIGURE 4.
A. Autocorrelation for a select item effect (β_4) from Gibbs Sampling. **B**. Greatly reduced autocorrelation from the inclusion of additional decorrelating Metropolis steps. The degree of autocorrelation is typical for all participant and item parameters.

alone; identifiability comes from the random effects assumptions expressed in the hierarchical prior. Even so, the constraint from the prior on the iteration-to-iteration values in the chains is small. The result is correlation. For example, when the α parameters are estimated high on average, the β parameters will shift as well. Likewise, a large estimate of θ attenuates the estimates of α and β.

Liu and Sabatti (2000) recommend adding additional Metropolis steps to decorrelate MCMC chains. For 2P-MAC, two decorrelating Metropolis steps are needed: (i) an additive step to decorrelate $\boldsymbol{\alpha}$ from $\boldsymbol{\beta}$, and (ii) a multiplicative step to decorrelate the sum of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from $\boldsymbol{\theta}$. The additive step is described in Morey et al. (2008), as it was necessary to analyze the 1P-MAC model. The multiplicative step needed for the 2P-MAC model may be useful to other researchers with similar models and autocorrelation problems. Hence, we describe it in some detail here. On every iteration of the chain, the Metropolis step proceeds as follows:

Step 1. Sample v from some proposal distribution restricted to positive values. Let $\pi(v)$ denote the density function of the proposal distribution. Selection of this distribution will be discussed later.

Step 2. Let $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\alpha}^{(t)}$, and $\boldsymbol{\beta}^{(t)}$ be the samples of $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ on iteration t of the Gibbs sampler. Let $\theta_i^{(t,v)} = \theta_i^{(t)}/v$, $\alpha_i^{(t,v)} = v\alpha_i^{(t)}$, and $\beta_j^{(t,v)} = v\beta_j^{(t)}$ on iteration t. These are candidates for acceptance.

Step 3. Evaluate u, the ratio of the posterior distributions for the candidate and the current sample:

$$u = \frac{p(\boldsymbol{\theta}^{(t,v)}, \boldsymbol{\alpha}^{(t,v)}, \boldsymbol{\beta}^{(t,v)}, \mathbf{w}, \sigma^2 \mid \mathbf{y})}{p(\boldsymbol{\theta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{w}, \sigma^2 \mid \mathbf{y})}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_\theta^2}\sum_{i=1}^{I}\left[\left(\theta_i^{(t)}/v\right)^2 - \left(\theta_i^{(t)}\right)^2\right]\right.\right.$$

$$+\sum_{i=1}^{I}\frac{1}{\sigma_\alpha^2}\left[\left(v\alpha_i^{(t)}\right)^2-\left(\alpha_i^{(t)}\right)^2\right]$$

$$+\sum_{j=1}^{J}\left[\left(v\beta_j^{(t)}\right)^2-\left(\beta_j^{(t)}\right)^2\right]\Bigg)\Bigg\}.$$

This term is a measure of how well the candidate values fit when compared with the samples from the tth iteration of the chain.

Step 4. Compute r:

$$r=\frac{v^J\pi(v^{-1})v^{-2}}{\pi(v)}.$$

This term ensures that the steps yield samples with the proper joint distribution.

Step 5. Accept candidate values $\boldsymbol{\theta}^{(t,v)}, \boldsymbol{\alpha}^{(t,v)}, \boldsymbol{\beta}^{(t,v)}$ with probability $\min(ur, 1)$.

A candidate distribution π must be specified. It is convenient if the expected value is about 1 to insure that both $v < 1$ and $v > 1$ are likely. Gamma distributions with equal rate and shape have this property. The variance of the proposal distribution may be tuned to provide for a desired acceptance rate in Step 5. We have found acceptance rates of between 0.25 and 0.5 provide for sufficient decorrelation.

Figure 4B shows a chain for a typical item parameter, β_4, with the inclusion of the additive decorrelating step described in Morey et al. (2008) and the multiplicative step described above. These steps do a good job of decorrelating the chains, leading to parameter estimates in minutes instead of hours.

4. Simulations

In order to assess the performance of the model, we performed simulations. The goal of these simulations is to determine whether model analysis can recover true parameters in reasonable sample sizes. In each of 500 simulations, 22 hypothetical participants judged 6 items 90 times each. These sample sizes are from the experiment discussed in Morey et al. (2008). True values of the six β parameters were (1.17, 0.35, −0.25, −0.91, −1.45, −1.89). True values for α_i and θ_i were drawn from independent Normal$(0, \sigma_\alpha^2 = .14)$ and $\text{TN}_{\Re^+}(0, \sigma_\theta^2 = .62)$ distributions, respectively. These true values come from the subsequent analysis of the Morey et al. data set.

For each of the 500 simulations, values of α_i and θ_i were drawn from their respective distributions. These values were then used to compute true latent scores x_{ij}. The true accuracies corresponding to these latent scores were used to generate binomial data for each participant-by-item combination. These simulated data were then analyzed via the 2P-MAC model, with the prior settings described previously. MCMC chains were run for 10,000 iterations, with the first 1,000 serving as burn-in.

Figure 5A provides a scatter plot of the true and estimated ability parameters. In total, there are $500 \times 22 = 11{,}000$ true participant abilities. Since a scatter plot of 11,000 points is to dense to be informative, a two-dimensional kernel density estate of the data is shown. Overall, the points cluster near the diagonal showing somewhat good recovery. There is a small degree of curvilinearity due to difficulty estimating participants who have extremely low latent ability (those 2 standard deviations below the mean). These participants are at chance for many of the items. Consequently, there is a paucity of information to discriminate how little ability they have. This paucity leads to a greater influence of the prior and subsequent shrinkage of the estimate. The

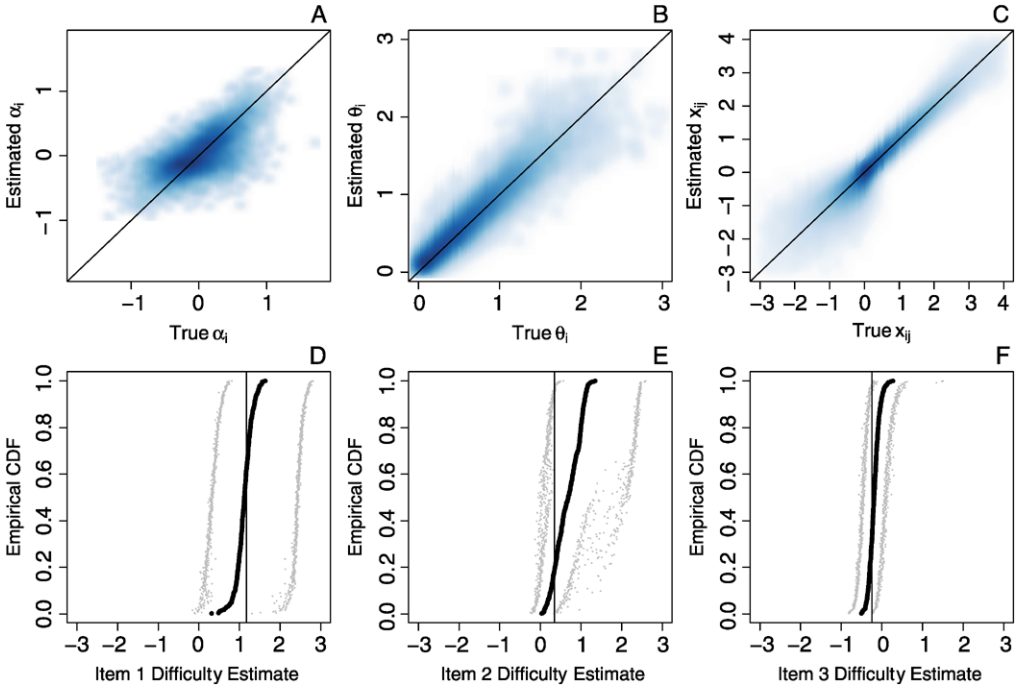

FIGURE 5.
Simulation results. **A–C**: Smooth scatterplots of estimated value against true value for the parameters α_i, θ_i, and $x_{ij} = \theta_i(\alpha_i - \beta_j)$, respectively. **D–F**: CDF coverage plots for model estimates of β_1, β_2, and β_3. *Dark points* represent point estimates (posterior mean); *light points* on either side of each dark point represent corresponding 95% posterior credible intervals. The *dark*, *solid vertical line* represents the true value.

effect is slight and occurs only for those simulated participants who had near baseline accuracies for all items. Whereas these participants inherently provide no information other than that they perform poorly, the shrinkage is not problematic. Figures 5B and C show the same plots for discriminability and true score. The points lie close to the diagonal, indicating good recovery of participant discriminability and latent score estimates.

Recovery of the true values of $\beta_1, \ldots, \beta_6$ is more nuanced and requires examination of the posterior distributions. Rather than show the posterior distributions for each β_i from each of the 500 simulations, we summarize the results with "CDF coverage plots" as follows. The dark dotted line is the empirical CDF of the β_i posterior-mean, point estimate. On either side of each dark point is a smaller, light point. The area between these two light points is the 95% credible interval that corresponds to the point estimate. Figure 5D–F show CDF coverage plots for parameters β_1, β_2, and β_3. The solid vertical line denotes the true value.

The recovery for the most difficult item (β_1, Figure 5D) is quite good, and this is surprising. In fact, the true value is so low that only 1 in 1,000 simulated participants have true performance above baseline. Hence, there is almost no information for the model to use, and recovery should be quite poor. The reason for the seemingly good recovery of β_1 is that the estimate of 1.17 is due solely to the prior. The standard normal prior on difficulties limits how large this estimate may be. In this case, for this true value and sample size, the shrinkage from the prior serendipitously leads serendipitously to good estimates. The lower bound of the credible interval does reflect information; the true value is probably greater than the bound. The upper bound and posterior mean, however, simply reflect the choice of the prior.

Figure 5E shows the CDF coverage plot for β_2, the second most difficult item. Inspection of the upper bound on difficulty reveals a mixture of two types of simulations. On some simulations,

no estimated abilities were above β_2, and the estimated value of β_2 reflected the prior. On other simulations, there were estimated abilities above β_2 and the estimate of β_2 reflects these values. This mixture leads to estimates of β_2 that are upwardly biased. The credible intervals, fortunately, quantify the amount of uncertainty well: over 95% of the credible intervals contain the true value.

Figure 5F shows the CDF coverage plot for parameter β_3. On almost all simulations, there were estimated abilities greater than this item's difficulty. Hence, there was much information to localize the parameter and the prior had almost no effect. Credible intervals are narrow and over 95% contain the true value. The average parameter estimate is close to the true value for the parameter. The remaining item difficulty parameters are estimated as well as β_3. In sum, the model estimates item difficulty well if at least one participant is estimated above chance for all items. Therefore, items that are so hard such that nobody does well are not useful for parameter estimation; this problem also applies to conventional IRT models.

5. Analysis of the Morey et al. (2008) Data Set

In the previous section, we noted the misspecification of 1P-MAC to the Morey et al. (2008) data set. In this section, we show the 2P-MAC analysis. The classification accuracy of each participant at each duration is shown in Figure 6A; each line represents the accuracy of a participant. Average performance is near chance for the most difficult durations, and increases for higher durations.

The data were analyzed with the 2P-MAC model, using the prior settings described previously. Chains were run for 10,000 iterations, and the first 1,000 iterations served as burn-in. The chains exhibited low autocorrelation and good convergence when decorrelating steps were included.[2] Figure 4B shows a typical chain for an item parameter (β_4).

Estimates of item difficulty parameters (β) and their 95% posterior credible intervals are shown as a function of duration in Figure 6B. Not only do these difficulties decrease with duration, as one would expect, there is a striking linear relationship between the logarithm of duration and difficulty. The line in Figure 6B shows the weighted least squares best fit. The linear fit is impressive precisely because the MAC models provide no a priori specification of the relationship between duration and difficulty—this relationship reflects underlying structure in the data. Figure 6C shows participant ability and discriminability as a scatter plot. There is no evident relationship, though this may not be so surprising with only 22 participants.

The linear relationship between log-duration and item difficulty aids in the estimate of a duration threshold for each participant. Using the parameter estimates from the weighted least squares regression, the threshold estimate, denoted $\hat{d}_i$, may be obtained by solving $\hat{\alpha}_i = 6.2 - 1.7 \log \hat{d}_i$. These values range from 25 ms to 54 ms as indicated on the right-hand axis of Figure 6C.

To assess model fit, standardized residuals r_{ij} may be computed as:

$$r_{ij} = \frac{y_{ij} - N_{ij}\hat{p}_{ij}}{\sqrt{N_{ij}\hat{p}_{ij}(1-\hat{p}_{ij})}}$$

where $\hat{p}_{ij}$ is $\Phi(\hat{x}_{ij} \vee 0)$ and $\hat{x}_{ij}$ is the posterior estimate of latent true score x_{ij}. If the model fits well, these residuals will fall between -1.96 and 1.96 for about 95% of participant-by-item combinations. Standardized residuals for the 2P-MAC fit are shown in Figure 6D as a function of

[2]To check convergence, we used Geweke's congergence statistic (Geweke, 1992). Of 52 parameters, 4 had Geweke statistics outside $(-1.96, 1.96)$, which is in line with what is expected assuming convergence. To be sure, we reran the analysis with 1,000,000 chain iterations. Parameter estimates did not change substantially.

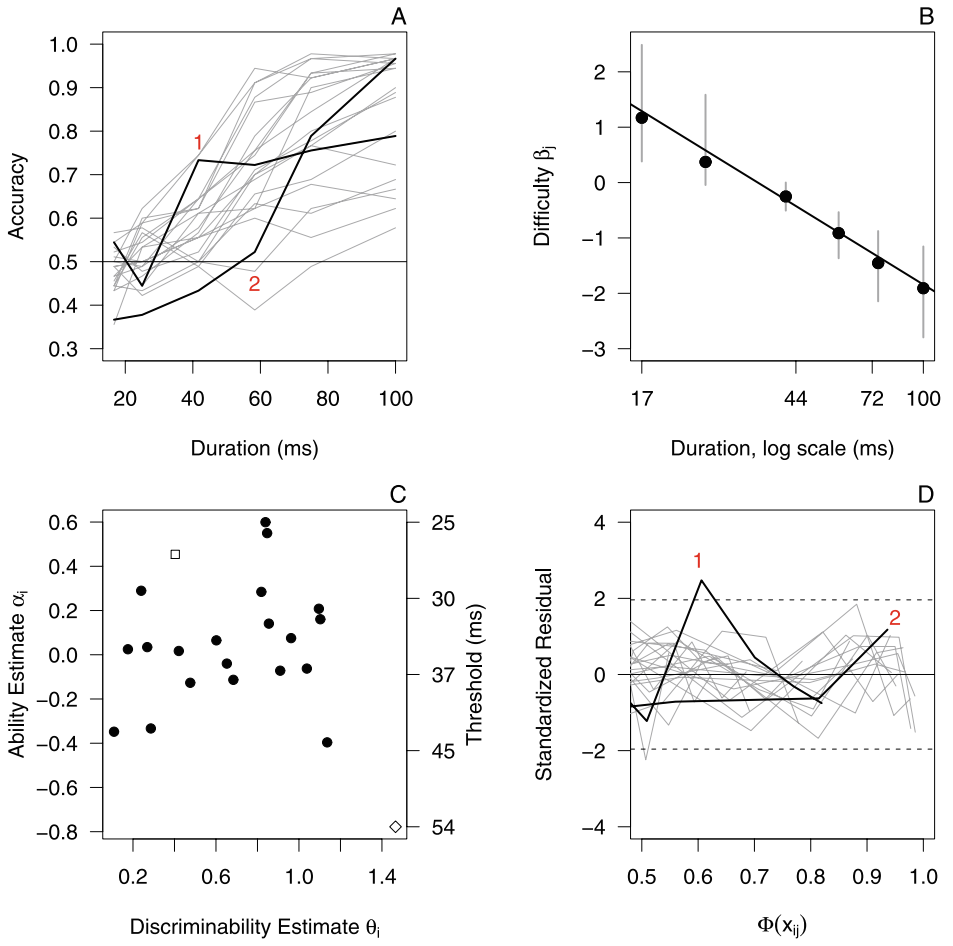

FIGURE 6.
Results of the described experiment and analyses. **A**: Accuracy as function of stimulus duration for all 22 participants. *Each line* represents a participant. Participants marked 'a' and 'b' are the same participants shown in Figure 3. **B**: 2P-MAC item difficulty (β) estimates as a function of duration. Error bars are 95% credible intervals. Fitted regression line is best linear fit with weighted least squares. **C**: Ability estimates (α) as a function of discriminability estimates (θ) for all 22 participants. The *square* and *diamond* are Participants 1 and 2 from Panel A, respectively. *Right axis* shows the threshold in ms corresponding to each ability. **D**: Standardized residuals from the 2P-MAC analysis. The *dark lines* correspond to participants whose residuals were large and patterned in the 1P-MAC analysis (see Figure 3B). These residuals are greatly improved.

predicted performance for above-baseline combinations. Each line represents a single participant. The residuals generally fall around zero and inside the 95% bounds, denoted by horizontal broken lines. No systematic trends are obvious. These residuals may be contrasted with those for the 1P-MAC model in Figure 3B. The fit of the 2P-MAC model greatly improves upon the 1P-MAC model; the highlighted participants "a" and "b" no longer show large, sweeping residuals. Of course, the demonstration of improved residuals with a more flexible model is not unexpected. To account for differences in flexibility, we also computed the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) for 1P-MAC and 2P-MAC. The DIC favored (was smaller for) 2P-MAC by 47, a sizable amount considering that DIC is on a logarithmic scale. Thus, the 2P-MAC model fits well and provides a substantial improvement over 1P-MAC for this set.

In summary, the 2P-MAC model fits revealed two interesting psychologically-substantive results. First, item difficulty appears linearly related to the logarithm of duration in the task described. Second, participants' abilities do not appear to be correlated to their discriminabilities. Exploring these potential psychophysical invariances is made possible in reasonable sample sizes using the 2P-MAC model.

6. Conclusions

The concept of a threshold is common in psychology. A stimulus is below threshold if the resulting performance on a measure is identical to an appropriate chance-level baseline. This notion may be implemented in IRT with a truncated normal link. We developed a two-parameter version using standard Bayesian techniques and applied the resulting model to a perceptual task.

Mass at Chance models fill an important gap in the psychophysics and psychometric literature. Many models fail to consider the possibility of either chance or ceiling levels of performance. For many applications, extreme performance levels may be uninteresting. However, in some research domains, establishing the extreme levels of performance is critical. As discussed previously, other techniques such as staircase procedures, confidence intervals on accuracy, or traditional IRT modeling fail to provide adequate solutions to characterizing performance at the extremes. MAC models explicitly include transitions from baseline levels of performance to higher performance. The 2P-MAC model developed here is a substantial improvement over the simpler versions of Rouder et al. (2007) and Morey et al. (2008).

We are optimistic that threshold links may be applicable in domains outside of subliminal priming. Consider the following two examples of assessing factual knowledge and assessing the effects of sleep deprivation on a cognitive skill. For the factual knowledge domain, there may be cases where the person has no knowledge whatsoever about the item. For instance, many Western 5th graders have no knowledge of the capital of China's Sichuan province. In these cases, the concept of a threshold is quite convenient—we simply model this item's difficulty as greater than the participants geographic ability. The resulting predicted performance is at the appropriate baseline. A second domain in which thresholds may be applicable is describing factors that adversely affect performance. Consider, for example, performance on cognitive tasks after a variable degree of sleep. We can define baseline performance as that following a normal 8-hour night's sleep. It is certainly true that many tasks exhibit declines after a very abbreviated nights sleep such as a 2-hour night's sleep. The question is whether there is some level that is equivalent to baseline: for example, a 7-hour night's sleep. That is, there may be a threshold on the length of sleep needed for baseline performance. If this threshold is not met, then performance declines. In sum, while the MAC models are motivated by psychophysics, they may be broadly applicable in many testing domains.

References

Abrams, R., Klinger, M., & Greenwald, A. (2002). Subliminal words activate semantic categories (not automated motor responses). *Psychonomic Bulletin and Review*, *9*, 100–106.

Albert, J.H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Andersen, E.B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Breitmeyer, B.G. (1984). *Visual masking: An integrative approach*. London: Oxford University Press.

Dehaene, S., Naccache, L., Le Clech, G., Koechlin, E., Mueller, M., & Dehaene-Lambertz, G. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597–600.

Eimer, M., & Schlaghecken, F. (2002). Links between conscious awareness and response inhibition: Evidence from masked priming. *Psychonomic Bulletin and Review*, *9*, 514–520.

Gelfand, A., & Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation. Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 4*. London: Oxford University Press.

Liu, S.J., & Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, *87*, 353–369.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Morey, R.D., Rouder, J.N., & Speckman, P.L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reingold, E.M., & Merikle, P.M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*, 563–575.

Rouder, J.N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573–604.

Rouder, J.N., Morey, R.D., Speckman, P.L., & Pratte, M.S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, *14*, 597–605.

Snodgrass, M., Bernat, E., & Shevrin, H. (2004). Unconscious perception: A model-based approach to method and evidence. *Perception & Psychophysics*, *66*, 888–895.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583–639.

Swaminathan, H., & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, *7*, 175–191.

Taylor, M.M., & Creelman, C.D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, *41*, 782–787.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701–1728.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time course for the visual perception and action priming. *Proceedings of the National Academy of Sciences*, *100*, 6275–6280.

Watson, A.B., & Pelli, D.G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113–120.

Manuscript Received: 14 APR 2008
Final Version Received: 10 MAR 2009
Published Online Date: 9 APR 2009